Developing Efficient Data Archive Designs for the State of Virginia

Simona Babiceanu
Brian Smith, Assistant Professor
Xiaoning Lu
Tim Ngov
Ramkumar Venkatanarayana

Presentation Outline

- STL (Smart Travel Lab)
- Why a database?
- ETL (Extraction, Loading and Transformation)
- MySQL staging database
- Oracle warehouse database
  - Hampton Roads Freeway
  - Northern Virginia Freeway
  - NVSTSS (Northern Virginia Smart Traffic Signal System)
- TMC Applications of Archived Data Operational Test
- Conclusions and Future Work

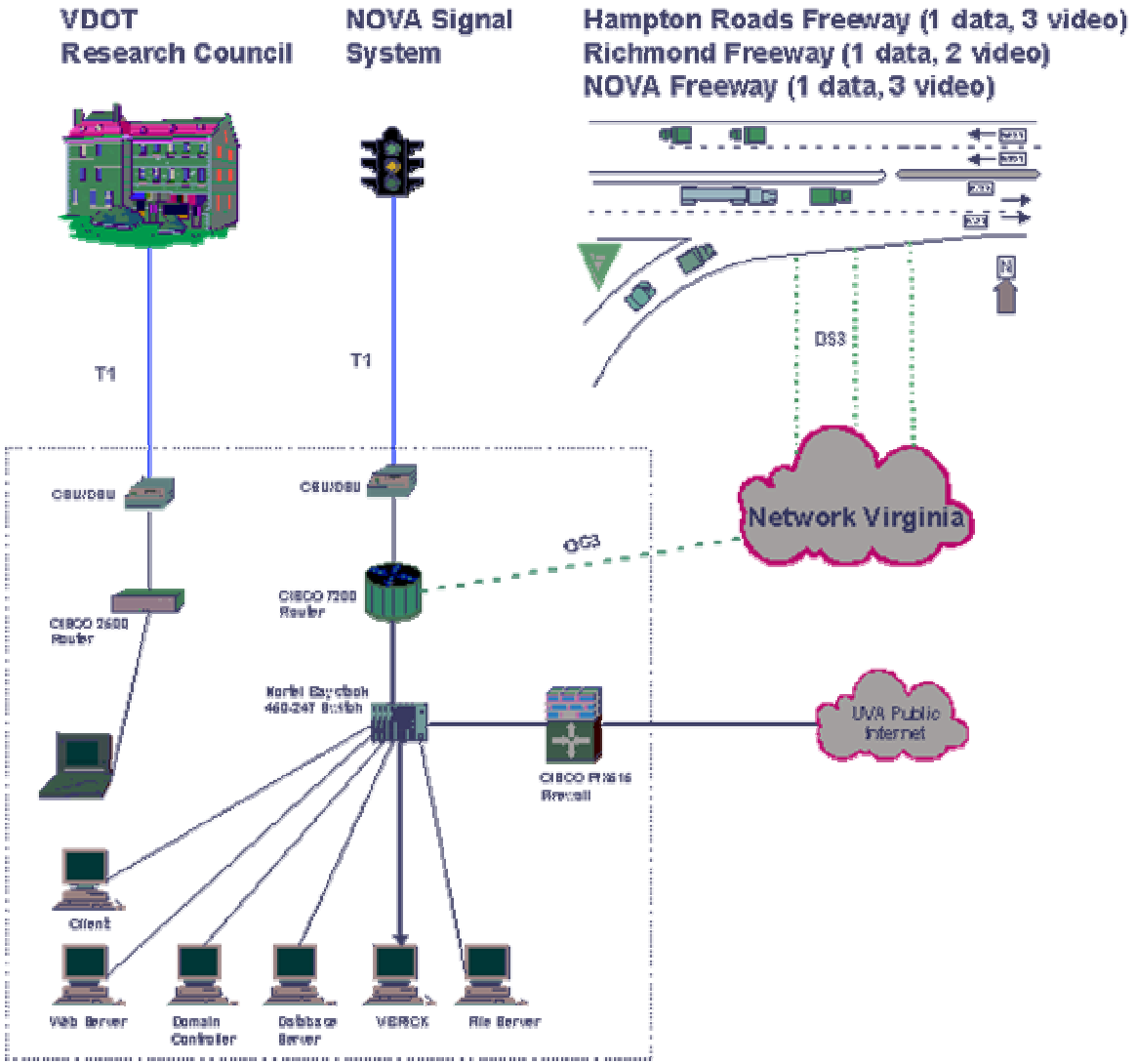
STL (Smart Travel Lab)

- State-of-art ITS research
- Joint effort between University of Virginia CTS (Center for Transportation Studies), CE Dept. and VTRC (Virginia Transportation Research Council), research branch of VDOT
- Designated official data archive repository of VDOT
- RT data streams
- Advanced IT

STL History

- Lab started in 1998
- Initially, archiving only HR freeway data
- Currently
  - HR – 1 freeway data, 3 video
  - Dump of HR Incident database 5.5 years
  - Richmond – 2 video channels
  - NOVA – 1 freeway data
  - NVSTSS – 1 arterial data
- 8 faculty, 3 researchers, 3 staff, 40+ students
- Smart Travel Van

Smart Travel Lab Network Connectivity



Why a Database?

- Relational database
- STL archives very large amounts of data every day
- Built-in search capabilities
- Easy to manage data (inserts, deletes, updates)
- Easy to build complex queries (SQL or GUI tool - Oracle Discoverer)
- Support to put data on the web
- Built-in authentication and authorization

ETL – Northern Virginia Freeway Data

- NOVA STC delivers detector data (speed, volume, occupancy, lane) as a flat text file to the lab via ftp every 10 seconds
- STL parses the 10-second files and aggregates them to 1-minute records that are inserted into a staging database
- Approx. 1000 detectors
- Approx. 1,5 million records/day for a full day

ETL – Hampton Roads Freeway Data

- HRSTC polls detectors every 20 seconds, aggregates data to 2 minutes, station level
- STL fetches the HRSTC station data (speed, volume, occupancy, lane) as a flat text file via ftp every 2 minutes
- STL parses the file and inserts the records into a staging database
- Currently 114 freeway stations
- Approx. 82,000 records/day for a full day

ETL – Data Validity Testing

- Percentage of bad data sometimes quite high
- 5 screening tests ([2])
- Abnormality value ([1])
  - Measures how close the data is to “normal” values for a particular day of week, time of day, station
  - Uses past behavior and compares current data with a historical average
- Data imputation – use a good “guessed” value instead of a “bad” value

MySQL Staging Database

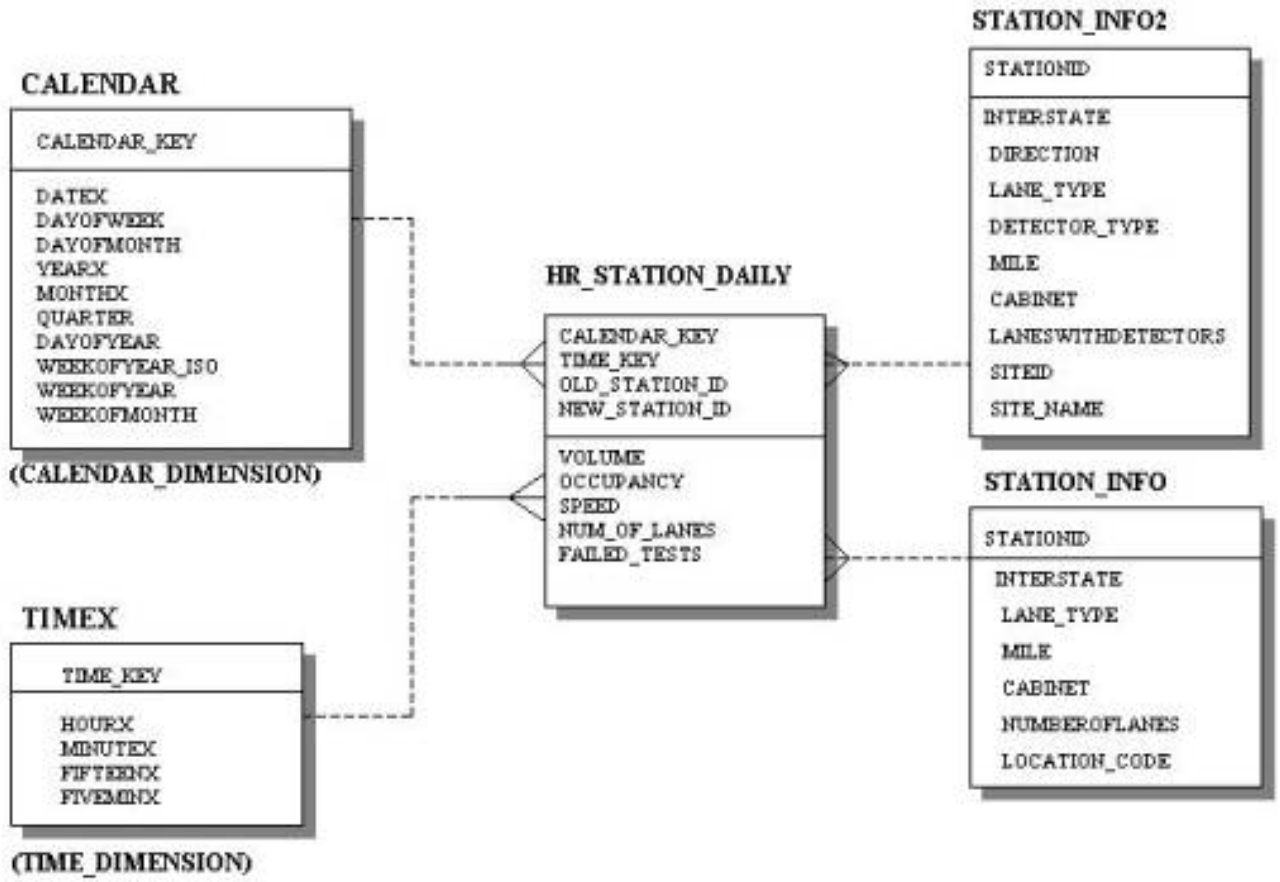
- Staging idea: ease some of the burden from main database
- Started using MySQL in summer 2002
- Chose it because
  - Planning on switching to MySQL replication to get data HRSTC data
  - Free, reliable, easy to use, has large user base
- STL parser uses JDBC to connect to MySQL
- Disadvantage: extra step

Oracle 8i Warehouse Database ([3])

- Warehouse – data repository geared towards optimizing searches (joins) and data analysis
  - Dimension model – *fact* (central) table containing keys into *dimension* tables and possibly other fields; simpler schema possibly at the expense of some data duplication
- Transactional database – geared towards optimizing inserts, deletes and updates (OLTP)
  - ER model – no restriction on topology; more complicated schema but space-efficient
- Nightly processes get yesterday's data from the staging database and insert it into the warehouse

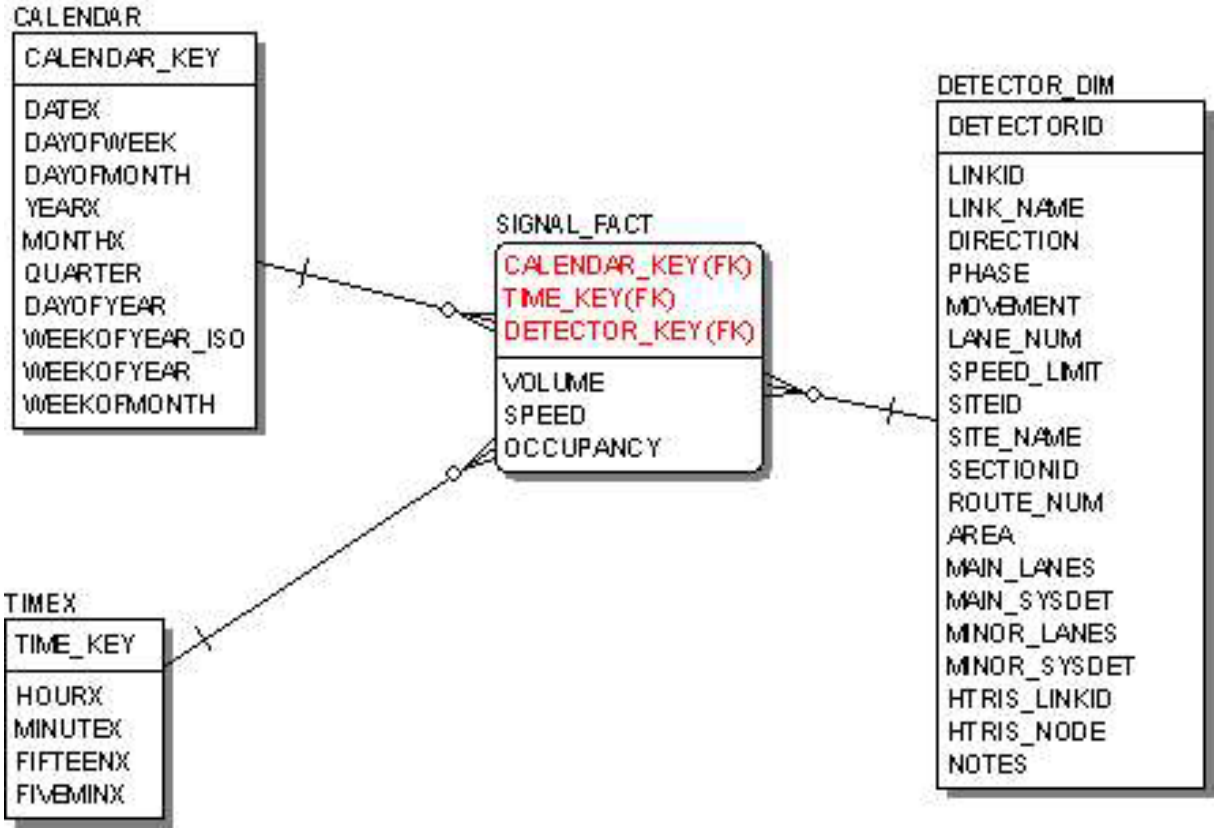
Smart Travel VIRGINIA

Hampton Roads New Station Data ER Diagram



NVSTSS

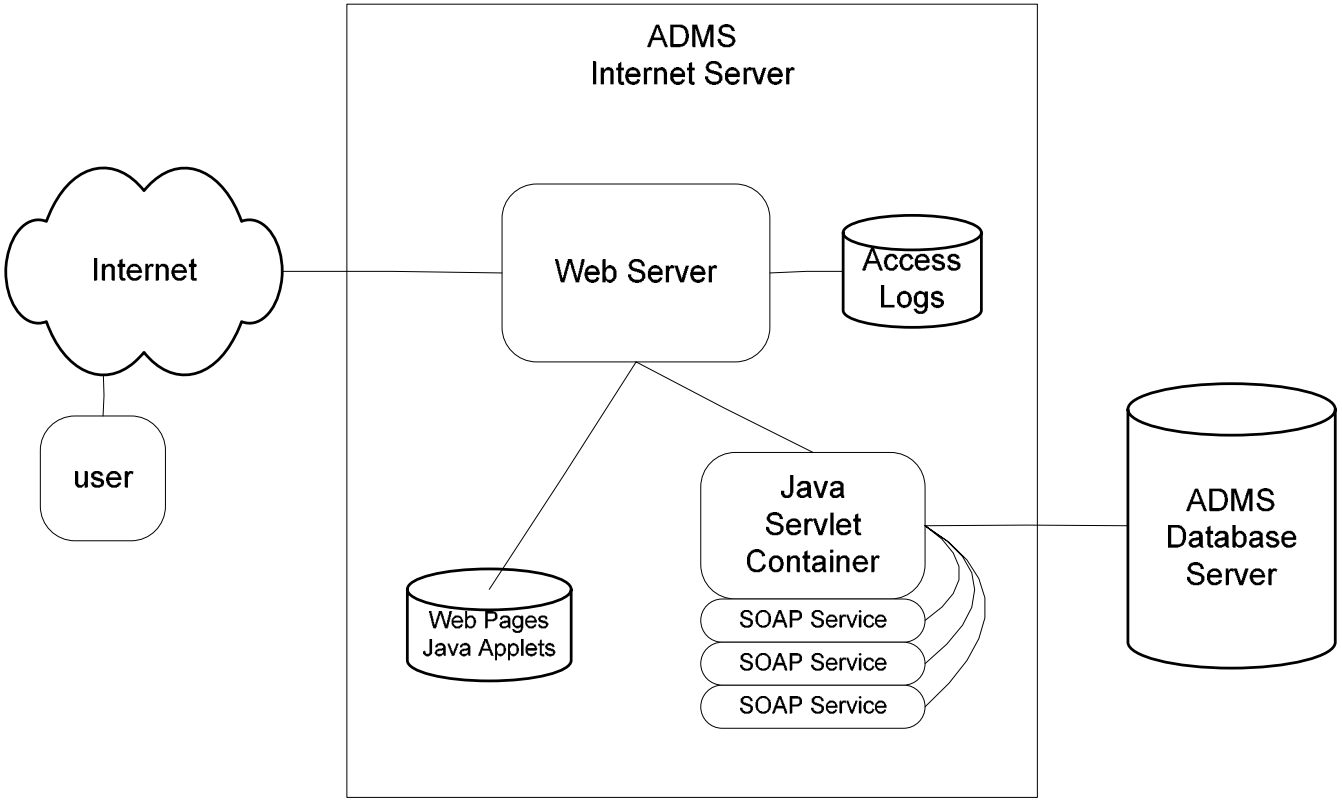
- Flat file every 15 mins, 1000+ traffic signal stations
- Approx. 144000 records/day, for a full day



TMC Applications of Archived Data Operational Test

- Monitor ITS data quality
- Detour/evacuation/construction/transit route planning/impact
- Real-time incident management support
- Historical data average output
- Access to classification data
- Quantitative impact of weather
- Regional planning support
- Support for 511/ATIS efforts
- Forecasting

TMC Applications of Archived Data Operational Test



Conclusions and Future Work

- Took advantage of database and warehouse features
- Star design offered better performance for complex queries
- Broke task into manageable parts

- Renewed need to have good data
  - Finish data imputation
- We would like to archive more data:
  - HR detector, ramp, tunnel, classification data
  - City of Norfolk green time and signal data
  - WIM truck data
  - Richmond STC data
  - real-time HR incident data
- Keep ETL, database design as flexible as possible
- Implementing standards mid-stream

Contacts

- STL http://smarttravellab.virginia.edu
- Brian L. Smith briansmith@virginia.edu
  (434) 243-8585
- Simona Babiceanu sbabiceanu@virginia.edu
- Xiaoning Lu xl4v@virginia.edu
- Tim Ngov timngov@virginia.edu
  (434) 924-4548
- Ram Venkatanarayana Ramkumar@virginia.edu
  (434) 293-1992
  Dept. of Civil Eng. University of Virginia
  P.O. Box 400742 Charlottesville VA 22904

Bibliography

- [1] Turochy, R. and Smith, B. “Alternative Approaches to Condition Monitoring in Freeway Management Systems”, VTRC 02-R8, January 2002
- [2] Turochy, R. and Smith, B. “A New Procedure For Detector Data Screening In Traffic Management Systems”, TRB 00-0842
- [3] Smith, B., Lewis, D. and Hammond, R. “Design of Archival Traffic Databases: A Quantitative Investigation Into the Application of Advanced Data Modeling Concepts”, TRB 03